# NAVAL POSTGRADUATE SCHOOL Monterey, California



## ALLOCATION OF JOBS TO UNEQUALLY-CAPABLE PROCESSORS: A PLANNING APPROACH

by

Donald P. Gaver / Patricia A. Jacobs / Kevin Becker / Siriphong Lawphongpanich /

September 1996

Approved for public release; distribution is unlimited.

Prepared for:

Naval Postgraduate School Monterey, CA 93943

FedDocs D 208.14/2 NPS-OR-96-010 Ted Lor 2202 14/2 DPS 08-90-010

## DUDLEY KNOX LIBRARY NAVAL POSTGRADUATE SCHOOL MONTEREY CA 93943-5101

#### NAVAL POSTGRADUATE SCHOOL MONTEREY, CA 93943-5000

Rear Admiral M. J. Evans Superintendent Richard Elster Provost

This report was prepared and funded by the Naval Postgraduate School.

Reproduction of all or part of this report is authorized.

This report was prepared by:



REPORT DOCUMENTATION PAGE				Form Approved OMB No. 0704-0188
Public reporting burden for this collection of it gathering and maintaining the data needed, collection of information, including suggestion Davis Highway, Suite 1204, Arlington, VA	arding this bur	rden estimate or any other aspect of this operations and Reports, 1215 Jetterson		
1. AGENCY USE ONLY (Leave blank)		3. REPORT TYPE AND DA		
	September 1996	Technical		
4. TITLE AND SUBTITLE Allocation of Jobs to U Approach	nequally-Capable Proce	ssors: A Planning	5. FUNDING	g numbers /A
6. AUTHOR(S) Donald P. Gaver, Patri Siriphong Lawphongp	cia A. Jacobs, Kevin Bec vanich	ker, and		
7. PERFORMING ORGANIZATION NAM	E(S) AND ADDRESS(ES)			MING ORGANIZATION
Naval Postgraduate So Monterey, CA 93943	chool			S-OR-96-010
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A				
11. SUPPLEMENTARY NOTES 12a. DISTRIBUTION / AVAILABILITY	STATEMENT		12b. DISTR	
Approved for public re				
13. ABSTRACT (Maximum 200 words) This paper addresses the problem in which jobs of different types arrive at a system that consists of a collection of individual and somewhat diverse processors. The processors differ in that each may specialize in one job type, but may also do others. Job types that are totally incompatible with a processor have an infinite service on that processor, but degrees of incompatibility may exist, and are modeled here. Using static queuing models, several practical performance measures may be evaluated, and optimal allocation of jobs to processors are obtained by solving linear and nonlinear programming problems. To illustrate, several numerical examples are provided. It is shown that jobs are not always most advantageously assigned to their most expert servers.				
14. SUBJECT TERMS queuing models; allocation of jobs to unequal servers				5. NUMBER OF PAGES 31 5. PRICE CODE
17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICAT OF ABSTRACT	ION 20	. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified		UL

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89) Prescribed by ANSI Std. 239-18



## ALLOCATION OF JOBS TO UNEQUALLY-CAPABLE PROCESSORS: A PLANNING APPROACH

D. P. Gaver P. A. Jacobs K. Becker S. Lawphongpanich

#### ABSTRACT

This paper addresses the problem in which jobs of different types arrive at a system that consists of a collection of individual and somewhat diverse processors. The processors differ in that each may specialize in one job type, but may also do others. Job types that are totally incompatible with a processor have an infinite service on that processor, but degrees of incompatibility may exist, and are modeled here. Using static queuing models, several practical performance measures may be evaluated, and optimal allocation of jobs to processors are obtained by solving linear and nonlinear programming problems. To illustrate, several numerical examples are provided. It is shown that jobs are not always most advantageously assigned to their most expert servers.

#### 1. Problem Formulation

Consider a service system whose individual service facilities (servers) are *unequal* in their capacities to serve different job types. This means that, in some sense, a job of type j is most expeditiously done by a server of type i = i(j), and quantifiably less so by other servers. It may well be that a good start is achieved by assigning jobs to those servers who require the least mean time to finish them,

but clearly this is not optimal if it tends to overload a few efficient servers and leaves others idle. A good dynamic approach might be to allocate the excess backlog of the efficient servers to others whenever that backlog is excessive. But this requires constant monitoring of queue lengths, or possibly current elapsed job service time. We do not investigate such dynamic rules here.

We propose to study a static queuing model that selects from the incoming job traffic stream of jobs of classes j = 1, 2, ..., J a subset that is directed to each of the available servers: if a job is of type j it goes to server type i with probability  $a_{ji}$ . The procedure is based on classical M/G/1 queuing theory, and requires mathematical programming in order to optimize the allocation from the perspective of either servers or jobs (customers). The allocation can be both deterministic or randomized. In the deterministic case, all jobs of one type are assigned to a single server. For the randomized allocation, if a job of type j arrives, it can be assigned to server i with probability  $a_{ji}$ .

#### 2. Model

The system studied is made up of I > 1 single servers with different capabilities. The system is confronted by a Poisson (rate  $\lambda$ ) rate of demand of jobs, but of different job types: with independent probability  $p_j$  an arrival is a job of type  $j, j \in (1, 2, ..., J)$ . The different capabilities of servers to handle (serve) jobs of different types is reflected in their service times: the service time of job j on server i, denoted by  $S_{ji}$ , may tend to vary systematically with i for any of many possible reasons, one being each of training or recent experience by i with job type j. In fact, for some job types j' and some servers i',  $S_{j'i'}$  is effectively infinite if the servers in question have no capability to handle those particular jobs, so  $E[S_{j'i'}] = \infty$ ; this is a case of total incompatibility, and certainly exists in many practical settings. On the other hand it may be necessary to allocate jobs to

servers with which they are somewhat incompatible in order to avoid overloading the more compatible servers. It is to be expected that this phenomenon may tend to occur more extensively as traffic intensity increases, e.g. if  $\lambda$  increases, in which case a good planning policy choice of  $a_{ji}$  may well stave off disaster: the need to hastily add servers or reject jobs.

To investigate the effect of cross-assignment we introduce a static set of assignment probabilities  $a_{ji}$ . Thus if a type-*j* job arrives in (t, t + dt) with probability  $\lambda p_j dt + o(dt)$ , it is assigned to server *i* with probability  $a_{ji}$ . We think of  $a_{ji}$  as a decision variable to be determined so as to optimize some measure of system performance. Total expected delay to all arriving jobs is one such measure, but the delays of some jobs may be more undesirable than those of others, in which case a total weighted expected delay can be studied. Note that, for job type *j*, we can select  $a_{ji}$  to be any real positive (actually non-negative) number such that  $\sum_{i=1}^{l} a_{ji} = 1$ , in which case a randomization device is needed to allocate an arrival of type *j* to its server, i.e.,  $a_{ji}$  represents a randomized job

allocation. A more easily implemented approach would be to choose a single "best" *i*-value, *i*(*j*) for each *j*, and set

$$a_{ji} = \begin{cases} 1 & \text{if } i = i(j) \\ 0 & \text{otherwise} \end{cases}$$

This approach is a deterministic allocation of jobs since it sends all jobs of type j to server i(j). Later, it is shown that, although easier to implement, deterministic job allocation may not be desirable in practice.

Both deterministic and randomized job allocations provide a simple independent Poisson stream of jobs with independent service times to the individual servers. Moreover, if jobs are treated in arrival order (first-come, firstserved) at all servers, the stationary delay experienced at each of the *I* servers can be calculated using known results, e.g. the Pollaczek-Khintchine-Kendall formula. Parenthetically, static priority rules can also be followed, and formulas for long-run results are already available, but this extension is postponed.

Clearly the arrival rate at server *i* of all job types is

$$\lambda_i = \lambda \sum_{j=1}^{J} p_j a_{ji} = \lambda \overline{p}_i \quad i = 1, 2, ..., I$$
(2.1)

and by standard results  $\lambda_i$  is the rate of a Poisson process independent of those prevailing for other servers ( $\neq i$ ). Now consider the effective service time,  $S_i$ , of jobs arriving at server *i*. A new job is of type *j* with probability  $p_j$ , and if it is dispatched to server *i* it experiences service time  $S_{ji}$ , so

$$S_i = S_{ji}$$
 with probability  $\frac{p_j a_{ji}}{\overline{p}_i}$ . (2.2)

Moments of  $S_i$  are calculated by conditioning, e.g.

$$E[S_i] = \sum_{j=1}^{J} E[S_{ji}] p_j a_{ji} / \overline{p}_i$$
(2.3)

$$E[S_i^2] = \sum_{j=1}^J E[S_{ji}^2] p_j a_{ji} / \overline{p}_i.$$
(2.4)

Also the Laplace transform is

$$E\left[e^{-\theta S_i}\right] = \sum_{j=1}^{J} E\left[e^{-\theta S_{ji}}\right] p_j a_{ji} / \overline{p}_i.$$
(2.5)

This latter becomes useful in case one wishes to use the probability of a long wait or delay as an optimization criterion; see Gaver and Jacobs (1988).

#### 3. Performance Measures

Several service-system/queuing performance measures are relevant to overall system performance and can be formally evaluated. Then mathematical programming techniques can be applied to obtain optimal  $\{a_{ij}\}$  values.

#### 3.1 Traffic Intensity Parameters and Associated Optimization Problems

It has already been stated that  $\lambda_i = \lambda \overline{p}_i$ , and that  $E[S_i] = \sum_{i=1}^{j} p_j a_{ji} E[S_{ji}] / \overline{p}_i$ , so

the *i*<sup>th</sup> server's traffic intensity is

$$\rho_i = \lambda_i E[S_i] = \lambda \overline{p}_i E[S_i] = \lambda \sum_{j=1}^J p_j a_{ji} E[S_{ji}].$$
(3.1)

Using (3.1), several optimization problems can be formulated from the perspective of the servers. The first problem is to find a job allocation that minimizes a linearly weighted combination of the traffic intensities, i.e., solve the following optimization problem.

P1:

$$\min_{a_{ji}} \quad \lambda \sum_{i=1}^{I} c_i \sum_{j=1}^{J} p_j a_{ji} E[S_{ji}]$$

sub

ject to 
$$\lambda \sum_{j=1}^{J} p_j a_{ji} E[S_{ji}] < 1 \forall i = 1, \dots I$$
 (3.2)

$$\sum_{i=1}^{J} a_{ji} = 1 \quad \forall j = 1, \dots J$$
 (3.3)

$$a_{ji} \ge 0 \quad \forall i = 1, \dots I; \ j = 1, \dots J$$
 (3.4)

where the first constraint, (3.2), restricts the traffic intensity for each server *i* to be less than one in order to ensure the existence of long-run stationary delays. The second constraint ensures that the probability that job type *j* is assigned to server

*i* sums to one. The parameter  $c_i$  in the objective function is the weight for the traffic intensity of server *i*. When  $c_i = 1$ , the objective function is equivalent to minimizing the average traffic intensity of the *I* servers.

When  $\lambda$  is sufficiently small, allocation  $a_{ji}$  satisfying constraints (3.3) and (3.4) may automatically satisfy constraint (3.2) and it may be eliminated. In this case, problem P1 has a simple solution. By interchanging the two summations, the objective function for P1 can be equivalently written as

$$\min \sum_{j=1}^{J} p_j \sum_{i=1}^{I} c_i a_{ij} E[S_{ji}].$$

For each job *j*, let  $\tilde{i}(j) = \arg \min\{c_i E[S_{ji}] : i = 1, ... I\}$  and allocate jobs as follows:

$$a_{ji}^{*} = \begin{cases} 1 & \text{if } i = \tilde{i}(j) \\ 0 & \text{otherwise} \end{cases} \quad \forall j = 1, \dots J.$$
(3.5)

The allocation in (3.5) clearly satisfies (3.3) and (3.4). Moreover, it is also assumed to satisfy (3.4) when  $\lambda$  is sufficiently small. By construction,  $\tilde{i}(j)$  is the least cost server and, when  $c_i = 1$  for all i,  $\tilde{i}(j)$  is the "best" or the most qualified server for job j. Since  $a_{ji}^*$  assigns job j to its least cost server, the corresponding objective function value must also be minimal. So the randomized and deterministic job allocations are the same when  $\lambda$  is sufficiently small.

When  $\lambda$  is too large, the allocation in (3.5) may not be feasible. In fact, there may not exist any feasible deterministic job allocation and the probabilistic job allocation may be the only choice. In which case, it is still optimal to assign the major proportion of job j to server  $\tilde{i}(j)$ . To minimize total cost, it is only necessary to divert just enough of jobs j to other servers to ensure that constraint (3.2) is satisfied. In practice, this may not be acceptable since server  $\tilde{i}(j)$  is likely

to have a traffic intensity (loading) close to one. To prevent this, several alternative optimization models for probabilistic job allocation are described below.

The first model minimizes the weighted sum of squared intensity of each server and can be stated as follows:

P2: 
$$\min \sum_{i=1}^{I} c_i \left( \lambda \sum_{j=1}^{J} p_j a_{ji} E[S_{ji}] \right)^2$$

subject to constraints (3.2), (3.3), and (3.4).

The squared objective function of P2 penalizes servers with higher traffic intensity more than those with lower intensities. P2 can be solved as a quadratic programming problem (see, e.g., Bazarra, Sherali and Shetty, 1993). Since the objective function for P2 is convex, the solution is guaranteed to be globally optimal.

The second model tends to equalize the traffic intensity among all servers by minimizing the maximum intensity, i.e.,

P3: 
$$\min \max_{i} \left\{ \lambda \sum_{j=1}^{J} p_{j} a_{ji} E[S_{ji}] \right\}$$

subject to constraints (3.2), (3.3), and (3.4).

The objective function for P3 is piecewise linear and convex. Thus, P3 is a nonlinear programming problem. However, it can be converted into a linear problem by introducing an auxiliary variable z and additional constraints to calculate the maximum traffic intensity as follows:

min z

subject to constraints (3.2), (3.3), (3.4), and

$$\lambda \sum_{j=1}^{J} p_j a_{ji} E[S_{ji}] \le z \quad \forall i = 1, \dots I$$
(3.6)

An optimal solution to P4 tends to assign jobs to servers so that their intensities all equal  $z^*$ , the optimal value of z.

Replacing constraint (3.4) with the following constraint translates randomized job allocation problems P2–P4 into deterministic ones.

$$a_{ii} \in \{0,1\} \quad \forall i = 1,...I; \ j = 1,...J.$$
 (3.7)

Constraint (3.7) simply restricts  $a_{ji}$  to be either 0 or 1 and the resulting problems become integer programming problems, a more difficult class of problems to solve. However, if the optimal allocation differs from the allocation in (3.5), then some jobs may not be given to the most qualified servers and the optimal allocation may be hard to accept in practice. For the remainder, we focus on the randomized job allocation problems with a more operationally meaningful objective (penalty) function based on expected or mean job delay, or mean nonlinearly-length-penalized job delay.

#### 3.2 Optimization of Mean Functions of Total Job Delay

The Pollaczek-Khintchine-Kendall formula provides the expected long-run waiting time E[W], at an M/G/1 system, so for server *i* we get

$$E[W_i] = \lambda_i E[S_i^2] \cdot \frac{1}{2(1-\rho_i)} \quad \text{if } \rho_i < 1.$$
(3.8)

A type-*j* job arrives in (t, t + dt) with probability  $\lambda p_j dt + o(dt)$ . With probability  $a_{ji}$  it is then assigned to server *i* where it experiences a total expected delay in system equal to  $E[W_i] + E[S_{ji}]$ . Thus the total expected delay for a job of type *j* is

$$E[D_{j}] = \sum_{i=1}^{I} a_{ji} \Big( E[W_{i}] + E[S_{ji}] \Big);$$
(3.9)

this becomes infinite if any  $\rho_i \ge 1$ . The total expected long-run *weighted* delay (per unit time) is

$$D(\underline{a}) = \sum_{j=1}^{J} \lambda p_j d_j E[D_j]$$

$$= \lambda \sum_{j=1}^{J} p_j d_j \left[ \sum_{i=1}^{I} a_{ji} \left( E[W_i] + E[S_{ji}] \right) \right].$$
(3.10)

From the perspective of jobs (or customers), it is natural to find a randomized job allocation that minimizes D(a) by solving the following nonlinear programming problem:

P5:

F

$$\min \sum_{j=1}^{I} p_j d_j \left[ \sum_{i=1}^{I} a_{ji} \left( E[W_i] + E[S_{ji}] \right) \right]$$
(3.11)

subject to constraints (3.2), (3.3), and (3.4)

where  $d_j$  is the weight for the delay of job j. When  $d_j = 1$ , the objective function of P5 is equivalent to minimizing the average delay of all J jobs. Alternately, the following problem minimizes the maximum expected delay:

P6: 
$$\min \max_{j} \left\{ \sum_{i=1}^{I} a_{ji} \left( E[W_i] + E[S_{ji}] \right) \right\}$$
(3.12)

subject to constraints (3.2), (3.3), and (3.4).

#### 3.3 Mean Non Linear Delay Penalty

Suppose it is important that the penalty for job delays be more stringent for long jobs, and in a manner that the (linear) long-run expectation of total delay does not reflect adequately. One such penalty parameterization is exponential: exact penalty  $E\left[e^{\theta_i\left(W_i+S_{ji}\right)}\right]$  where  $\theta_i > 0$ . The value of  $\theta_i$  is a decision maker's

choice. Classical M/G/1 theory says that the limiting transform is

$$E\left[e^{\theta_{i}W_{i}}\right] = \frac{1-\rho_{i}}{1-\rho_{i}\left\{\frac{E\left[e^{\theta_{i}S_{i}}\right]-1}{\theta_{i}E\left[S_{i}\right]}\right\}}$$
(3.13)

provided the denominator is positive. To satisfy this requirement, the rate of input to server i must be, in general, smaller than what is allowed by the expected long-run waiting time formula in (3.8). Of course,

$$E\left[e^{\theta_i S_i}\right] = \sum_{j=1}^J p_j a_{ji} E\left[e^{\theta_i S_{ji}}\right] / \sum_{j=1}^J p_j a_{ji}$$
(3.14)

can quickly grow large, or become formally infinite, if any assignments of job type j' (denoting members of a subset of all jobs) are submitted to server i. Analogous to P5, the problem of minimize weighted penalty can be written as follows:

 $\min \sum_{j=1}^{J} p_j d_j \left\{ \sum_{i=1}^{I} a_{ji} E\left[e^{\theta_i W_i}\right] E\left[e^{\theta_i S_{ji}}\right] \right\}$ 

P7

subject to constraints (3.2), (3.3), (3.4), and

$$0 < \rho_i \left\{ \frac{E[e^{\theta_i S_i}] - 1}{\theta_i E[S_i]} \right\} < 1 \quad \forall i = 1, 2, \dots I.$$

$$(3.15)$$

The additional constraint is to ensure that the expression for  $E[e^{\theta_i W_i}]$  in the objective function is well defined. As before, when  $d_j = 1$ , P7 simply minimizes the average delay penalty for all *J* jobs. Similarly, the problem of minimizing the maximum penalty can be written as

P8

$$\min\max_{j} \left\{ \sum_{i=1}^{I} a_{ji} E\left[e^{\theta_{i} W_{i}}\right] E\left[e^{\theta_{i} S_{ji}}\right] \right\}$$

subject to the constraints of (3.2), (3.3), (3.4), and (3.15).

#### 4. Incompatibility Models

Among all possible job to server assignments, it is convenient to stipulate that the assignment of job j to server j (or, equivalently, job i to server i) is the most compatible, i.e.,  $E[S_{ii}] \leq E[S_{ji}]$  and  $E[S_{ii}^2] \leq E[S_{ji}^2]$  for  $j \neq i$ . Below are two incompatibility models that fit this stipulation.

#### 4.1 Proportionality Model

Suppose we represent incompatibility simply as follows: for  $k_{ji} \ge 1$ , we assume  $S_{ji} = k_{ji}S_i$ , so

$$E[S_{ji}] = k_{ji} E[S_{ii}]$$
(4.1a)

$$E\left[S_{ji}^{2}\right] = k_{ji}^{2} E\left[S_{ii}^{2}\right]$$
(4.1b)

for  $\forall j \neq i$ . Clearly the inequalities noted above hold, and the greater k is made the more flagrant is the incompatibility.

#### 4.2 Random Interruption Model

An alternative and physically plausible model is as follows. Let  $v_{ji}$  denote a Poisson rate of interruptions incurred by a job of type *j* when processed on server type *i*. These interruptions occur when the server must consult for advice, look for necessary materials, or rectify a breakdown that occurs. It is assumed here that the job is not displaced from the server's attention (set aside) during the interruption, so no other jobs may be done during the interruption. Analysis of a set-aside option will be conducted later. Let  $I_{ji}$  denote the random duration of an

interruption; successive interruptions are independent. Let  $S_{ji}^{\#}$  be the effective service time for server *i* to complete a job of type *j*, including the interruptions that occur. Then

$$S_{ji}^{\#} = S_{ii} + \sum_{\ell=1}^{N(S_{ii})} I_{ji}(\ell)$$
(4.2)

where, given  $S_{ii}$ ,  $N(S_{ii})$  is Poisson ( $v_{ji}S_{ii}$ ).

By straightforward conditioning we find

$$E[S_{ji}^{\#}] = E[S_{ii}](1 + v_{ji}E[I_{ji}])$$
(4.3)

$$E\left[\left(S_{ji}^{\#}\right)^{2}\right] = E\left[S_{ii}^{2}\right]\left(1 + v_{ji}E\left[I_{ji}\right]\right)^{2} + v_{ji}E\left[S_{ii}\right]E\left[I_{ji}^{2}\right]$$
(4.4)

#### 5. Numerical Examples

The problems presented above, P1, P2, P4, P5, P6, P7, and P8 were implemented and solved using the General Algebraic Modeling System (GAMS) developed by Brooke, Kendrick and Meeraus (1992). (Recall that P4 is linear version of P3.) Among all the constraints in these problems, constraints (3.2) and (3.15) cannot be implemented on a finite precision computer. Our implementation replaces equations (3.2) and (3.15) by the following:

$$\lambda \sum_{j=1}^{J} p_{j} a_{ji} E[S_{ji}] \le 0.99 \quad \forall i = 1, \dots I.$$
(5.1)

$$.01 \le \rho_i \left\{ \frac{E[e^{\theta_i S_i}] - 1}{\theta_i E[S_i]} \right\} \le .99.$$
(5.2)

For P1, P2, P4, P5, and P6, our example assumes that the first and second moments for service time are determined by the random interruption model.

Results for the proportionality model are similar. The data for our example are as follows:

- i) I = J = 6,
- ii) The weight for the traffic intensity at server i,  $c_i$ , equals 1 for all i,
- iii) The cost of delay for job j,  $d_j$ , equals 1 for all j,
- iv) The probability of an incoming job being of type j,  $p_j$ , is as follows:

Job	1	2	3	4	5	6
p <sub>j</sub>	0.29	0.24	0.19	0.14	0.10	0.04

- v)  $v_{ji} = \text{Round}(U(1,4))$ , where U(a, b) is a Uniform random variable with parameters a and b, and Round() rounds a number to the nearest integer.
- vi)  $E[I_{ji}] = U(0.5, 1.5)$  and  $E[I_{ji}^2] = 2E[I_{ji}]^2$ .

vii) 
$$E[S_{ii}] = U(0.5, 1.0)$$
 and  $E[S_{ii}^2] = 2E[S_{ii}]^2$ .

Table 1 shows the resulting matrices for the first and second moments generated based on the data from v, vi and vii.

			$E[S_{ji}]$			
Jobl Job2 Job3 Job4 Job5 Job6	Serverl 0.5364 1.3794 2.4959 1.6340 1.7820 1.5059	Server2 1.3056 0.5878 1.6771 1.2613 2.0643 1.9296	Server3 3.8214 2.4318 0.7628 2.3811 1.6229 3.1550	Server4 2.6083 5.1551 3.7252 0.8751 1.5611 2.2705	Server5 2.2060 1.2537 1.5549 2.8333 0.5891 2.1116	Server6 1.3083 2.0121 0.9507 1.3421 1.4267 0.5171
			$E\left[S_{ji}^2\right]$			
Jobl Job2	Server1 0.5754 5.1302	Server2 4.2853 0.6911	<b>Server3</b> 35.3372 15.4788	<b>Server4</b> 15.8944 63.6172	<b>Server5</b> 14.1712 4.6431	Server6 4.6341 10.9792
Job3 Job4 Job5 Job6	16.0385 6.8372 9.2434 5.7039	6.9712 4.7252 12.2305 10.5089	1.1638 14.7727 7.2066 24.9085	32.3951 1.5316 5.9493 11.7939	5.8913 21.7556 0.6940 12.8533	2.5347 4.9190 5.1381 0.5347

Table 1: The first and second	l moments of the time server	<i>i</i> takes to perform job <i>j</i> .
-------------------------------	------------------------------	--

Using the above data, the maximum job arrival rate,  $\lambda_{max}$  can be obtained by solving the following linear programming problem:

P9:

#### $\lambda_{\max} = \max \lambda$

subject to constraints (3.3), (3.4), and

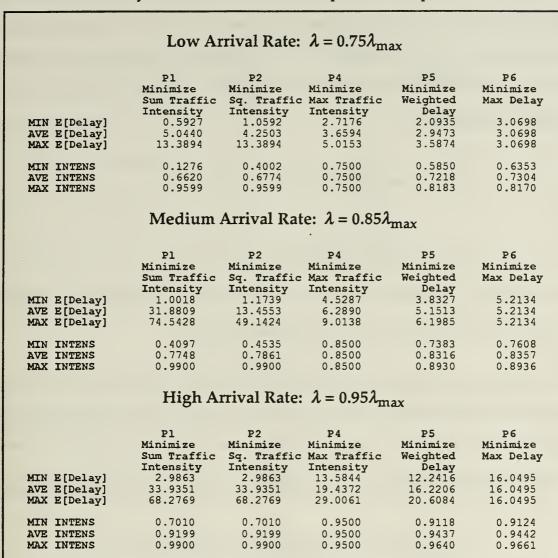
$$\lambda \sum_{j=1}^{J} p_j a_{ji} E[S_{ji}] \le 1 \quad \forall i = 1, \dots I.$$
(5.3)

This problem yields  $\lambda_{max} = 8.23$ . Corresponding to this  $\lambda_{max}$ , there is also a feasible allocation,  $\hat{a}_{ii}$ , such that

$$\lambda_{\max} \sum_{j=1}^{J} p_j \hat{a}_{ji} E[S_{ji}] = 1 \quad \forall i = 1, \dots I.$$
(5.4)

Table 2 summarizes the results for three different arrival rates: low  $(0.75\lambda_{max})$ , medium  $(0.85\lambda_{max})$  and high  $(0.95\lambda_{max})$ . The first two allocations mimimize the sum (P1) and the squared traffic intensities (P2) yield higher delays than the other three allocations. On the other hand, P1 and P2 produce smaller traffic intensities on the average. However, the maximum intensity for P1 and P2 for all three arrival rates are close to 1. Except for the low arrival rate, the maximum traffic intensity for both P1 and P2 is 0.99, the maximum allowed by constraint (5.1). This may not be desirable in practice.

The last three allocations minimize the maximum traffic intensity (P4), weighted delays (P5) and maximum delay (P6), have similar average intensity. Since minimizing the maximum traffic intensity does not directly take into account the delay, P4 generates slightly higher job delays than the other two. Observe that the traffic intensity for P4 is the same for every server. Similarly, the expected delay for every job is the same for problem P6.



#### Table 2: Summary of results from various optimization problems

It is interesting to note that problems P5 and P6 which minimize functions of job delays produce allocations with traffic intensities similar to those generated by P4 which minimizes the maximum intensity. This similarity is probably due to the fact that, from (3.8), lower traffic intensity implies shorter waiting time which, in turn, implies shorter delay via (3.9). Moreover, this similarity illustrates that planning from the two perspectives, servers or jobs (customers), does not have to always be conflicting.

Problem P4, minimizing the maximum traffic intensity, tends to equalize the intensity at all servers. In Table 2, the minimum, average, and maximum values of traffic intensity for P4 all equal , if  $\lambda = \alpha \lambda_{max}$ . In fact, the solution to problem P9, i.e.,  $\hat{a}_{ji}$ , also solves P4 because  $\hat{a}_{ji}$  is a feasible allocation and multiplying (5.3) by  $\alpha$  yields the following

$$\alpha \lambda_{\max} \sum_{j=1}^{J} p_j \hat{a}_{ji} E[S_{ji}] = \alpha \qquad \forall i = 1, \dots, I$$
$$\lambda \sum_{j=1}^{J} p_j \hat{a}_{ji} E[S_{ji}] = \alpha \qquad \forall i = 1, \dots, I.$$

So, the allocation  $\hat{a}_{ji}$  produces the same traffic intensity,  $\alpha$ , at all servers. Since  $\alpha$  is the maximum intensity,  $\hat{a}_{ji}$  must be optimal to P4 also.

Unlike the other problems, the distribution for service times must be specified for problems P7 and P8. To simplify our illustration, replace v to vii with the following:

viii) The service time  $S_{ii}$  has a gamma distribution with parameters  $\alpha_i$  and  $\beta_i$ , and

$$E[S_{ii}] = \beta_i / \alpha_i$$
 and  $E[S_{ii}^2] = (\beta_i^2 + \beta_i) / \alpha_i^2$ , and

ix) The service time  $S_{ji}$  has a gamma distribution with parameters  $\alpha_i$  and  $k_{ji}\beta_i$ , and

$$E[S_{ji}] = k_{ji}\beta_i/\alpha_i$$
 and  $E[S_{ji}^2] = \left(\left(k_{ji}\beta_i\right)^2 + k_{ji}\beta_i\right)/\alpha_i^2$ .

The above assumptions is similar to the proportionality model, in that equation (4.1a) holds. However, equation (4.1b) does not and the following holds instead:

$$E\left[S_{ji}^2\right] < k_{ji}^2 E\left[S_{ii}^2\right] \qquad \forall j \neq i.$$

$$E\left[e^{\theta_i S_{ji}}\right] = \left(1 - \frac{\theta_i Var\left[S_{ji}\right]}{E\left[S_{ji}\right]}\right)^{-E\left[S_{ji}\right]^2 / Var\left[S_{ji}\right]}$$

 $\forall i \text{ and } j.$ 

For our example,

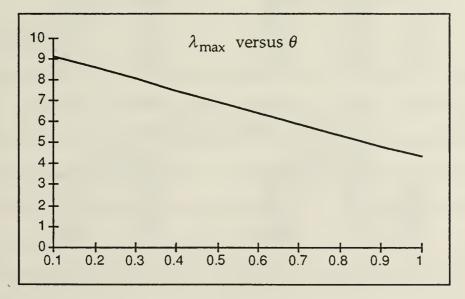
1) 
$$\alpha_i = 1 \text{ and } \beta_i = 2 \forall i$$

2) 
$$k_{ji} = \frac{|j-i|}{2} + 1, \quad \forall j \neq i.$$

Determining the value of  $\lambda_{max}$  now requires solving problem P9 with the addition of the following constraint:

$$0 \le \rho_i \left\{ \frac{E\left[e^{\theta_i S_i}\right] - 1}{\theta_i E[S_i]} \right\} \le 1, \qquad \forall i.$$
(5.5)

Since (5.5) depends on  $\theta$ , improper choice for  $\theta$  may render problems P7 and P8 infeasible. Figure 1 below displays a graph of  $\lambda_{max}$  as a function of  $\theta$  under the assumption that  $\theta_i = \theta$  for all *i*. So, the choice of  $\theta$  limits the application of



**Figure 1**: A graph of  $\lambda_{max}$  as a function of  $\theta$  under the assumption that  $\theta_i = \theta$  for all *i*.

problem P7 and P8. Certainly, when the arrival rate is 0.7, it is not possible to choose  $\theta$  to be 1 and vice versa. Tables 3 and 4 compare the results with  $\theta = 0.7$  for two pairs of problems, P5 & P7 and P6 & P8, respectively. In Table 3, minimizing weighted job delays (P5) and job penalties (P7) generate similar job delays for all three arrival rates. On the other hand, only the job penalties for low arrival rate are similar. At the high arrival rate, the infinite penalties for jobs 1 and 6 under P5 indicate that the penalty function is undefined, i.e., constraint (3.15) is not satisfied. Problems P6 and P8 display a similar behavior in Table 4.

Table 3: Comparing job delays and penalties for problems P5 and P7.

	Low Arrival I	Rate: $\lambda = 0.75 \lambda_{max}$	
Job 3 Job 4 Job 5	P7: Minimize           Weighted Penalty           DELAY         PENALT           1.1957         4.364           0.0157         3.401           0.9296         2.885           0.8693         2.716           0.8134         2.579           0.7664         2.605	Y         DELAY           9         1.2562           3         1.0504           0         0.8663           7         0.7537           9         0.7008	ed Delay PENALTY 6.0654 3.7571 2.5424 2.1365 2.0217
Ave	1.0008 3.372	8 0.9873	3.7493
	Medium Arriva	l Rate: $\lambda = 0.85 \lambda_{\rm m}$	ax
Job 2 Job 3 Job 4 Job 5	P7: Minimize           Weighted Penalty           DELAY         PENALT           1.3395         6.463           1.1525         4.969           1.0767         4.132           1.0059         3.848           0.9456         3.705           0.8858         3.845           1.1404         4.915           High Arrival         1.12	Y         DELAY           5         1.4309           5         1.1788           0         0.9700           6         0.8595           6         0.8320           0         0.9493	ed Delay PENALTY 14.1940 5.4984 3.1213 2.5732 2.6513 5.1916 6.8620
Job 2 Job 3 Job 4 Job 5 Job 6	P7: Minimize           Weighted Penalty           DELAY         PENALT           1.5106         15.554           1.3238         11.591           1.2673         9.383           1.917         8.821           1.1185         8.655           1.0326         9.275           1.3166         11.546	P5: Min Weight D 1.6392 0 1.3401 0 1.1036 2 0.9955 0 0.9977 8 1.1772	nimize ed Delay PENALTY ∞ 14.0168 4.4200 3.4966 4.3668 ∞

			-		
	Low	Arrival Rate:	2 - 0.752		
	LOW P	Anivai Kate:	$\lambda = 0.75 \Lambda_{\text{max}}$		
	P8: Mini	mize	P6: Minir	nize	
	Maximum	Penalty	Maximum De	elav	
	DELAY	PENALTY	DELAY	PENALTY	
Job 1		3 7031			
Job 2	1.0730	3.7831 3.7831 3.7831	1.0998 1.0998	3.8304	
Job 3	1.0803	3.7831	1.0998	3.7653	
Job 4	1.0803	3./831	1.0998	4.0381	
Job 5	1.0209	3.7831	1.0998	4.0422	
Job 6	0.8637	3.0756	1.0025	3.8652	
Ave	1.0729	3.7548	1.0959	3.8667	
	Modium	Arrival Rate	$\therefore \lambda = 0.85 \lambda_{\rm max}$		
	wieurum	Allival Nate	$n = 0.05 n_{\rm max}$	c	
	P8: Mini	mize	P6: Minin		
	Maximum	Penalty	Maximum De	elay	
1	DELAY	PENALTY	DELAI	PERMITI	
Job 1	1.2847	5.6349 5.6349	1.2438	5.4339	
Job 2	1.2045	5.6349	1.2438		
Job 3	1.1525 1.0736	5.6349 4.9570	1.2438	8.7200	
Job 4	1.0736	4.9570	1.2438	12.7694	
Job 5	1.1845 1.0599	5.6349	1.2438	13.7092	
Job 6	1.0599	5.2289	1.2438	21.1060	
Ave	1.1918	5.5237	1.2438	8.7634	
	High A	Arrival Rate:	$\lambda = 0.95\lambda_{max}$		
	-				
	P8: Mini		P6: Minir	uze	
	Maximum	Penalty	Maximum De	elay	
	DELAY	PENALTY	DELAI	PENALII	
Job 1	1.4910	13.1304	1.4212	15.6351	
Job 2		13.1304 13.1304 11.9743	1.4212	6.7326	
Job 3	1.3079	13.1304	1.4212 1.4212	15.5092	
JOD 4	1 1217	11.9743 9.6943 7.9914		⊥-2211 ∞	
JOD 5	1.131/	9.6943	1.4212 1.2959	8	
JOD 6	1.0162	7.9914	1.2909	00	
Ave	1.3267	12.4194	1.4162	N/A	

Table 4: Comparing job delays and penalties for problems P6 and P8.

### 6. Application

One application of the optimization problems is in quantifying the benefit of additional training. In all of the above examples,  $p_6 = 0.04$ . So, there are not many jobs of type 6 in the system. Therefore, server 6, who is the expert for job type 6, has to process other types of jobs in order to, e.g., minimize the maximum expected delay among all 6 job types, (problem P6). In order to improve this performance measure, it is logical to train server 6 to become an expert at processing another job besides job 6. Assume that, if server 6 is trained to process job *i*, then

$$E[S_{6i}] = E[S_{ii}] \text{ and } E[S_{6i}^2] = E[S_{ii}^2].$$

By resolving problem P6 with the first two moments of its service times suitably modified, the training benefit for server 6 can be measured quantitatively. Using data from the interruption model above (i.e., i - vii) and high arrival rate, Figure 2 displays the maximum delay for each training alternative.

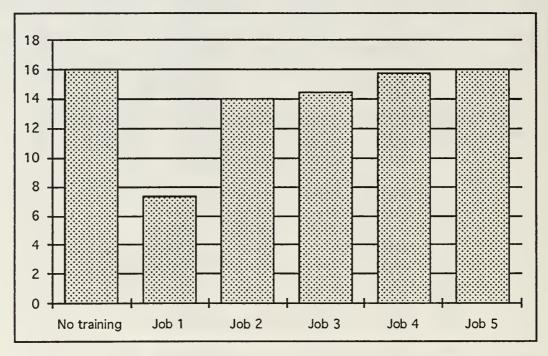


Figure 2: Maximum Job Delay vs. Training Alternatives for Server 6

As expected, it is best to train server 6 to process job 1, for doing so decreases the maximum job delay by the largest amount.

In general, it is also possible to formulate an optimization problem for assigning servers to training programs for a given budget constraint.

#### REFERENCES

- M.S. Bazaraa, H.D. Sherali, and C.M. Shetty, *Nonlinear Programming: Theory and Algorithms*, John Wiley & Sons, Inc., New York, New York (1993).
- A. Brooke, D. Kendrick, and A. Meeraus, *GAMS: A User's Guide*, Release 2.25, The Scientific Press, South San Francisco, California (1992).
- D.P. Gaver and P.A. Jacobs "Nonparametric estimation of the probability of a long delay in the M/G/1 queue," J. Royal Statist. Soc. Ser B., 5 (1988), 392-402.
- L. Kleinrock *Queueing Systems, Volume I: Theory,* John Wiley & Sons, New York, NY 1975.

٠.,

### DISTRIBUTION LIST

1.	Research Office (Code 09)
2.	Dudley Knox Library (Code 013)
3.	Defense Technical Information Center
4.	Therese Bilodeau
5.	Prof. Donald P. Gaver (Code OR/Gv)
6.	Prof. Patricia A. Jacobs (Code OR/Jc)
7.	Prof. Siriphong Lawphongpanich (Code OR/Lp)
8.	Mr. Kevin Becker
9.	Dr. J. Abrahams
10.	Dr. John C. Bailar

11.	Prof. D. R. Barr
12.	Dr. David Burman
13.	Center for Naval Analyses
14.	Dr. Edward G. Coffman, Jr
15.	Prof. Sir David Cox
16.	Prof. H. G. Daellenbach
17.	Dr. D. F. Daley
18.	Mr. Mike Davis
19.	Dr. B. Doshi

20.	Dr. Naihua Duan
21.	Dr. Guy Fayolle
22.	Dr. M. J. Fischer
23.	Prof. George S. Fishman
24.	Dr. Neil Gerr
25.	Dr. R. J. Gibbens
26.	Prof. Peter Glynn
27.	Prof. Linda V. Green
28.	Dr. Shlomo Halfin

29.	Prof. J. Michael Harrison
30.	Dr. P. Heidelberger
31.	Prof. D. L. Iglehart
32.	Institute for Defense Analysis
33.	Dr. F. P. Kelly
34.	LCDR Paul Knechtges
35.	Mr. Koh Peng Kong
36.	Prof. Guy Latouche
37.	Dr. A. J. Lawrance

38.	Prof. J. Lehoczky
39.	CAPT (Sel) Mike Lilienthal
40.	Dr. D. M. Lucantoni
41.	Dr. James R. Maar
42.	Dr. Colin Mallows
43.	Prof. M. Mazumdar
44.	Dr. James McKenna
45.	Prof. Paul Moose
46.	Dr. John A. Morrison

47.	Dr. T. J. Ott
48.	Dr. V. Ramaswami
49.	Dr. Martin Reiman
50.	Prof. Maria Rieders
51.	Dr. Rhonda Righter
52.	Prof. G. Shanthikumar
53.	Prof. N. D. Singpurwalla
54.	Prof. H. Solomon

.

55.	Prof. L. C. Thomas
56.	Dr. D. Vere-Jones
57.	Mr. Lai Kah Wah
58.	Dr. Ed Wegman
59.	Dr. L. Wein
60.	Dr. Alan Weiss
61.	Dr. P. Welch
62.	Prof. Roy Welsch



.